International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 2(3) (2026)

ijerst.editor@gmail.com

editor@ijerst.com

ISSN 2319-5991 www.ijerst.com
© 2026 IJERST. All Rights Reserved
Peer Reviewed, Referred & Indexed Journal

Research Paper

Hybrid Neuro-Acoustic Transductive Modeling of Infant Cry State Spaces for Early Physiological Condition Attribution

SK. Mahaboob Basha[1], Sapavath Tharun Nayak[1], Myadam Manish[1], M. Naga Srimaan[1], Pathulothu Laxman[1]

[1]Department of Computer Science and Engineering, [1]Sree Dattha Institute of Engineering and Science, Nagarjuna Sagar Road, Sheriguda, Ibrahimpatnam, Rangareddy Dist, 501510, Telangana, India.

ABSTRACT

The rapid advancement of intelligent audio-based monitoring systems has increased the demand for automated analysis of acoustic signals in healthcare-related applications, particularly in infant cry analysis where crying serves as the primary mode of communication. Accurately interpreting these signals is challenging due to variations in pitch, intensity, duration, and environmental noise, making reliable classification essential for effective assistance and early-stage monitoring. The problem lies in developing a system capable of distinguishing different categories of baby cries from unstructured audio data, where the existing system is primarily manual and depends on human observation and experience, leading to subjectivity, inconsistency, and lack of real-time applicability. These limitations highlight the need for an intelligent, automated, and scalable solution that minimizes human intervention while improving prediction reliability. To address this, the proposed system utilizes Mel-Frequency Cepstral Coefficients (MFCC) for feature extraction combined with multiple Machine Learning (ML) models, including Support Vector Machine (SVM), K-Nearest Neighbour (KNN), Decision Tree Classifier (DTC), Adaptive Boosting Classifier (ABC), and Linear Discriminant Analysis (LDA) for comparative evaluation, along with a Convolutional Neural Network (CNN) as the primary model for enhanced performance. The CNN enables the system to learn deeper feature representations from audio inputs, improving classification accuracy. The significance of this work lies in providing an automated and reliable method for baby cry classification, assisting caregivers in understanding infant needs, and demonstrating the effectiveness of integrating MFCC features with ML and CNN models for practical healthcare applications.

Key words: Infant Cry Classification, Audio Signal Processing, Machine Learning (ML), Deep Learning (DL), Mel-Frequency Cepstral Coefficients (MFCC), Healthcare Monitoring Systems.

1. INTRODUCTION

Infants depend entirely on crying as their innate and primary form of communication to express fundamental needs such as hunger, pain, discomfort, fatigue, or emotional distress. Since they lack the ability to articulate these needs verbally, accurate interpretation of crying patterns becomes essential for ensuring proper caregiving, emotional bonding, and healthy cognitive development during early life stages [1]. However, understanding the underlying cause of a baby's cry is inherently complex, as cry signals vary in pitch, intensity, duration, and temporal structure, often influenced by environmental noise and individual differences.

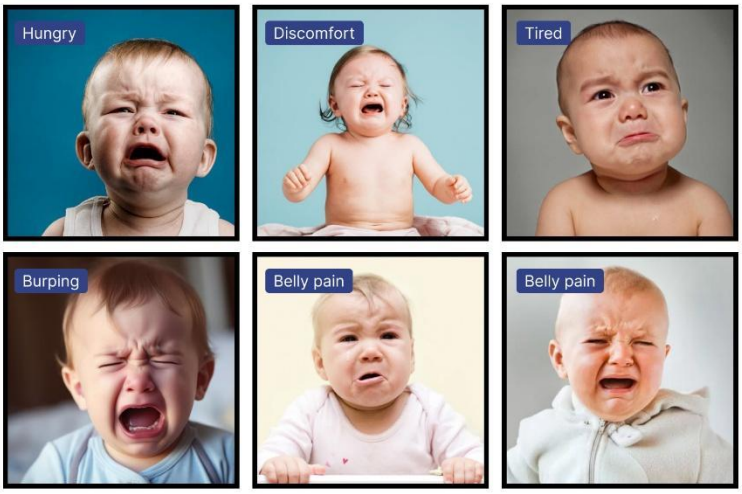

Fig. 1: Baby cry symptoms analysis.

Inexperienced parents, new caregivers, and even some healthcare professionals may struggle to correctly interpret these variations, leading to delayed or inappropriate responses. Such misinterpretations can result in prolonged crying episodes, increased stress levels for both the infant and caregiver, and potential negative effects on the child's emotional and neurological development [2].

To address these challenges, significant research efforts have been directed toward the automatic analysis and classification of infant crying sounds using computational approaches. This domain integrates digital signal processing techniques with Machine Learning (ML) algorithms to transform raw audio signals into meaningful representations and classify them into predefined categories. Typically, audio signals are preprocessed to remove noise and then converted into feature representations such as Mel-Frequency Cepstral Coefficients (MFCC), which effectively capture the spectral and perceptual characteristics of sound. These features are subsequently used as inputs for classification models that learn to distinguish between different types of cries based on patterns present in the data. Over time, various studies have explored different combinations of datasets, feature extraction techniques, and classification algorithms, forming a comprehensive foundation for automated cry analysis systems.

Given that infants rely solely on auditory signals to communicate, caregivers are often forced to make subjective judgments, which may not always be accurate [3]. This uncertainty can lead to inefficient caregiving decisions, causing frustration and anxiety for parents while prolonging discomfort for the infant. In response, modern technological advancements aim to provide intelligent and data-driven solutions that enhance the accuracy and consistency of cry interpretation [4]. By leveraging acoustic feature analysis and advanced learning models, these systems can identify subtle differences in cry patterns associated with specific needs. This enables more timely and appropriate responses, ultimately improving infant care and reducing caregiver stress. Furthermore, such systems demonstrate the potential to be integrated into real-world healthcare and monitoring applications, offering scalable and reliable support for caregivers and medical professionals alike [5].

2. LITERATURE SURVEY

Li, Y et al. [6] proposed a hybrid deep learning architecture that integrates Residual Networks (ResNet) with transformer-based mechanisms for infant cry classification. Their approach utilizes engineered Mel-Frequency Cepstral Coefficients (MFCC) features to represent acoustic patterns and incorporates Squeeze-and-Excitation (SE) attention modules within residual blocks to dynamically recalibrate channel-wise feature importance. This design enhances the model's ability to focus on relevant acoustic characteristics while suppressing noise. The proposed system demonstrated improved classification accuracy, faster convergence, and reduced training time compared to conventional ML approaches,

making it efficient and scalable for real-world applications. Mekhfioui, A et al. [7] developed a real-time embedded system for analyzing infant cry signals using both prosodic and cepstral features, particularly MFCC coefficients. The system captures live audio input through a microphone and processes it using a trained ML model to predict the underlying cause of crying, such as hunger, discomfort, or pain. The predicted outcome is displayed on a connected screen, and the system is further enhanced by transmitting collected data to a remote web server for storage and analysis. This integration enables continuous monitoring and supports large-scale data-driven improvements in classification performance. Zayed, H et al. [8] introduced a multimodal diagnostic framework aimed at identifying medical conditions in infants through cry signal analysis. Their system combines multiple feature domains, including harmonic ratio (HR) as a prosodic feature, Gammatone Frequency Cepstral Coefficients (GFCC) as a cepstral feature, and spectrogram-based image features extracted using a pretrained Convolutional Neural Network (CNN). By fusing these diverse features, the model captures both temporal and spectral characteristics of cry signals. The framework is specifically designed to detect critical conditions such as neonatal respiratory distress syndrome (RDS) and sepsis, achieving improved classification accuracy through multi-domain feature integration.

Liang, X et al. [9] proposed a hybrid deep learning model that combines CNN and Long Short-Term Memory (LSTM) networks to recognize various infant needs, including hunger, diaper change, emotional attention, and medical discomfort. The system uses MFCC features extracted from 1607 audio recordings of infant cries, each with a duration of 10 seconds. The CNN component captures spatial feature representations, while the LSTM component models temporal dependencies in the audio signals. The performance of this hybrid model was compared with a traditional Artificial Neural Network (ANN), demonstrating superior accuracy and better generalization capability. Herlea, A et al. [10] conducted a comprehensive evaluation of advanced deep learning architectures such as ResNet and EfficientNet for detecting infant crying sounds. Their study focuses on comparing the effectiveness of different input representations, including spectrograms and MFCC features, to determine which combination yields the best performance. The results highlight the importance of selecting appropriate feature representations and model architectures for improving the accuracy and reliability of baby monitoring systems. This work contributes to enhancing automated cry detection systems and supports the development of intelligent caregiving technologies. Khalilzad, N et al. [11] investigated the use of infant cry signals as biomarkers for distinguishing between neonatal pathological conditions, specifically sepsis and neonatal respiratory distress syndrome (RDS). Their approach combines features derived from harmonic ratio and Gammatone Frequency Cepstral Coefficients (GFCC) to capture both spectral and temporal characteristics of the audio signal. Machine Learning models such as Multilayer Perceptron and Support Vector Machine (SVM) were employed, with hyperparameters carefully tuned to optimize performance. The fusion of multiple feature sets significantly improved classification accuracy, achieving results above 95%, demonstrating the effectiveness of cry-based diagnosis.

Shayegh, M et al. [12] investigated the potential of self-supervised learning techniques for analyzing infant cry signals, focusing particularly on expiratory segments, which are considered highly informative for medical diagnosis. Their approach utilized advanced models such as wav2vec 2.0, WavLM, and HuBERT to automatically learn feature representations directly from raw audio data, eliminating the need for manual feature engineering. A simple fully connected classification layer was then applied to categorize infants into healthy, sepsis, or RDS groups. This method demonstrated strong capability in capturing complex acoustic patterns and improving classification accuracy while reducing preprocessing complexity. Tsalera, E et al. [13] explored the application of transfer learning for sound classification by retraining pretrained Convolutional Neural Network (CNN) models originally designed for image and audio tasks. Models such as GoogLeNet, SqueezeNet, ShuffleNet, VGGish, and YAMNet were evaluated to determine their adaptability to infant cry classification. The study examined the influence of key training parameters, including learning rate, batch size, optimizer

selection, and number of epochs, on both classification performance and computational efficiency. The results highlighted the effectiveness of transfer learning in reducing training time while maintaining high accuracy. Ozcan, H et al. [14] proposed a structured classification pipeline that integrates data augmentation, feature extraction, and Artificial Neural Network (ANN) modeling. The system employed MFCC features to represent audio signals based on human auditory perception, ensuring meaningful feature extraction. To enhance model performance, hyperparameter optimization was conducted using grid search techniques, allowing the selection of optimal configurations for improved generalization. The combination of augmentation and optimization significantly contributed to better classification accuracy. Khalilzad, N et al. [15] developed a classification framework for detecting infant sepsis using multiple acoustic feature sets, including MFCC, Spectral Entropy Cepstral Coefficients (SENCC), and Spectral Centroid Cepstral Coefficients (SCCC). These features were used as inputs to Machine Learning models such as K-Nearest Neighbour (KNN) and Support Vector Machine (SVM). The study also incorporated Bayesian Hyperparameter Optimization to fine-tune the models for each dataset. Performance was evaluated using metrics such as accuracy, F1-score, and Matthews Correlation Coefficient, demonstrating improved detection capability across different cry signal datasets.

3. PROPOSED SYSTEM

The proposed system implements a comprehensive pipeline for automated baby cry classification, integrating audio signal processing, ML techniques, and a CNN-based DL model within a unified framework. Initially, raw audio signals are provided as input and processed using Librosa, where MFCC features are extracted to capture perceptually relevant spectral characteristics of infant cries. These features are aggregated into numerical vectors and stored as structured datasets to enable efficient reuse and avoid repeated computation. The preprocessing stage ensures consistency in feature dimensions, removes noise-related variations, and prepares the data for reliable training. Subsequently, the dataset is divided into training and testing subsets using an 80/20 split to ensure unbiased evaluation. As illustrated in Fig. 2, the system incorporates multiple ML classifiers including SVM, KNN, DTC, ABC, and LDA, which act as baseline models for comparative performance analysis.

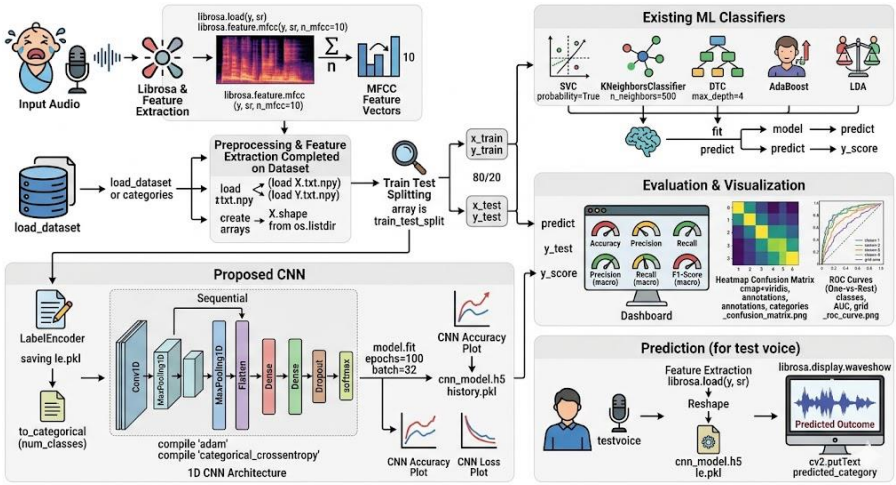

Fig. 2: Proposed system architecture

In parallel, a CNN model is designed to learn hierarchical representations from MFCC feature vectors, consisting of convolutional layers for feature extraction, pooling layers for dimensionality reduction, and dense layers for classification. The CNN model is trained using categorical cross-entropy loss and optimized with adaptive learning techniques, enabling it to capture complex temporal and spectral dependencies in audio signals. The system further evaluates all models using performance metrics such as accuracy, precision, recall, and F-score, along with visualization tools including confusion matrix and ROC curves to assess classification reliability. A GUI is integrated to facilitate user interaction,

allowing dataset upload, model training, and prediction execution. Finally, for unseen audio inputs, the trained CNN model performs classification into categories such as hungry, discomfort, burping, tired, and belly pain, with outputs visualized through waveform plots annotated with predicted labels. This end-to-end architecture ensures scalability, robustness, and practical applicability in real-time infant monitoring scenarios.

4. RESULTS DESCRIPTION

The Fig. 3. shows MFCC feature extraction interface confirms the successful preprocessing of the baby cry audio dataset, where the system automatically extracts MFCC from all input audio samples. After loading the dataset, the application identifies five cry categories such as belly pain, burping, discomfort, hungry, and tired and completes MFCC extraction, resulting in a structured feature matrix of size (1858, 10). This indicates that each audio sample is transformed into a compact 10-dimensional MFCC feature vector, making the data suitable for efficient training and evaluation of both traditional machine learning classifiers and the proposed CNN model for early baby health monitoring.

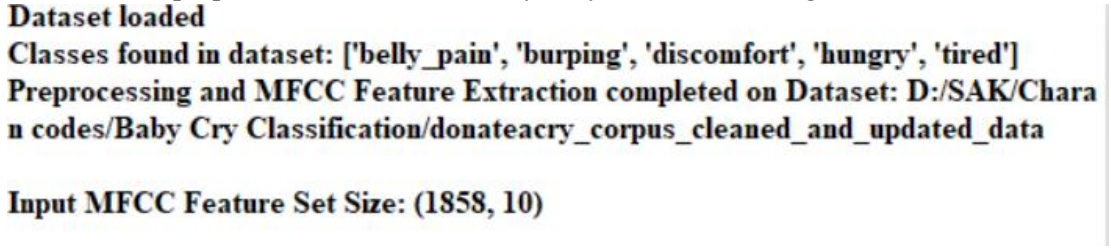
Dataset loaded
Classes found in dataset: ['belly_pain', 'burping', 'discomfort', 'hungry', 'tired']
Preprocessing and MFCC Feature Extraction completed on Dataset: D:/SAK/Chara
n codes/Baby Cry Classification/donateacry_corpus_cleaned_and_updated_data
Input MFCC Feature Set Size: (1858, 10)

Fig. 3. MFCC feature extraction completed

The Fig. 4. shows proposed CNN classifier results that clearly demonstrate a substantial improvement in baby cry classification performance compared to all traditional machine learning models. The confusion matrix shows an almost perfectly diagonal structure, indicating that the CNN accurately identifies each cry category—belly pain, burping, discomfort, hungry, and tired with minimal misclassification, as most samples are correctly mapped to their respective classes. This highlights the CNN's ability to learn discriminative temporal and spectral patterns directly from MFCC feature sequences rather than relying on handcrafted decision boundaries. The ROC curves further validate this superiority, with AUC values reaching 1.00 for four classes and 0.98 for the hungry class, and all curves positioned far above the random baseline. This reflects excellent class separability, strong probability estimation, and high robustness in multiclass prediction. Overall, the results confirm that the CNN model effectively captures the complex and nonlinear acoustic characteristics of baby cries, making it highly suitable for accurate early health monitoring and real-world deployment.

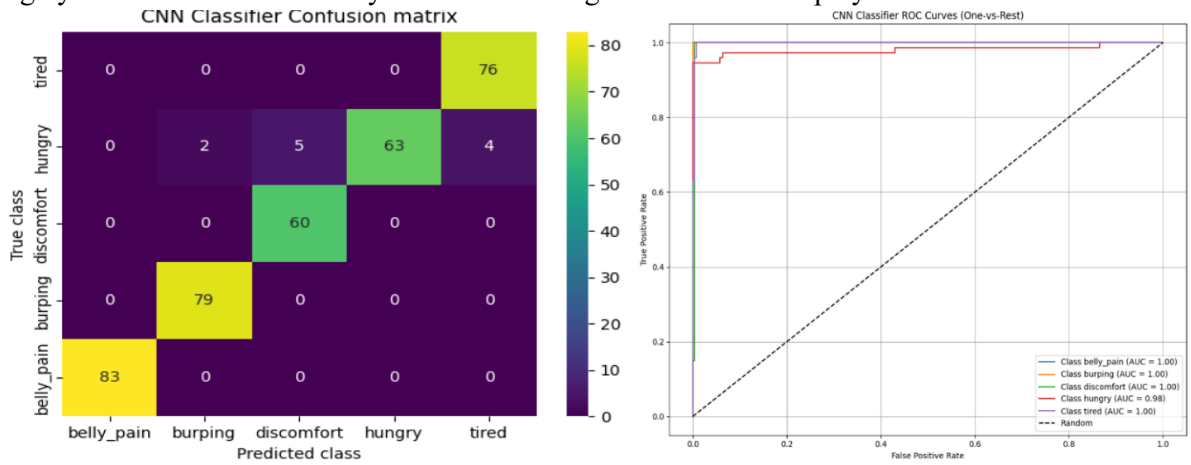

Fig. 4. Confusion matrix and ROC obtained using Proposed CNN

The Fig. 5. shows test audio prediction output using the proposed CNN model demonstrates the system's ability to accurately classify an unseen baby cry signal. After extracting MFCC features from the input audio, the trained CNN model analyzes the temporal and spectral patterns and predicts the cry

category as "tired," which is clearly displayed on the waveform visualization. The amplitude–time plot provides an intuitive representation of the cry signal while simultaneously confirming the model's decision, thereby validating the effectiveness of the proposed CNN for real-time baby cry classification and early health monitoring.

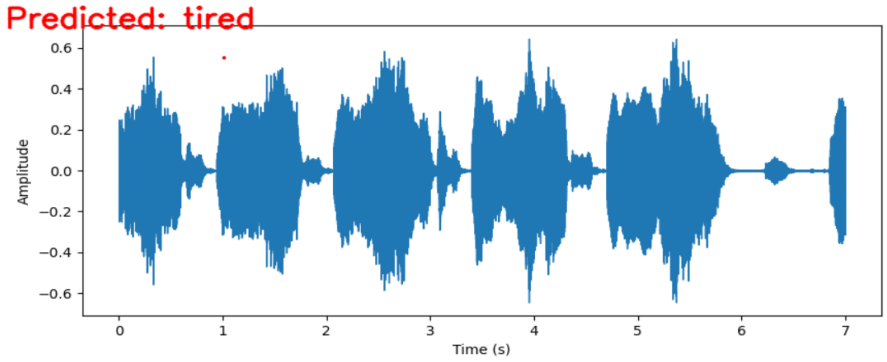

Fig. 5. Prediction on test sample obtained using Proposed CNN

The Fig. 6. depicts the prediction result on the test audio sample using the proposed CNN model shows that the system has correctly identified the baby cry category as "burping." After extracting MFCC features from the input audio signal, the CNN effectively captures distinctive temporal and spectral patterns associated with burping cries. The displayed waveform, along with the predicted label, confirms the model's strong classification capability and demonstrates its effectiveness for accurate and real-time baby cry analysis in early health monitoring applications.

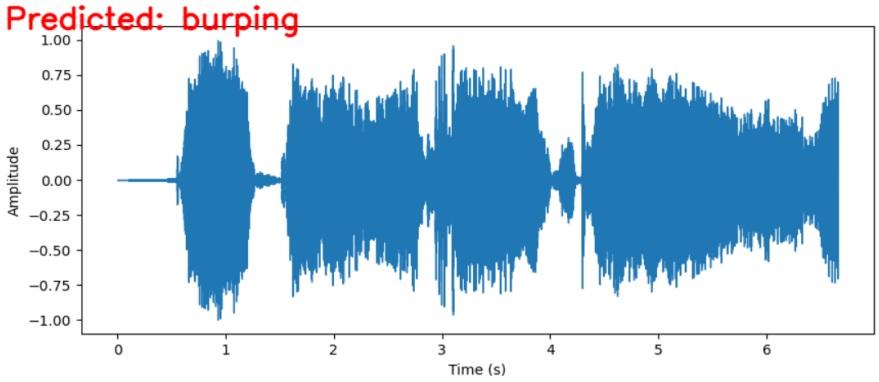

Fig. 6: Prediction on test sample obtained using Proposed CNN

Table 1: Performance comparison for the SVM, DTC, KNN, ADB, LDA and Proposed CNN Model

Algorithms Name	Accuracy	Precision	Recall	F-score
SVM	46.50%	51.65%	46.10%	47.03%
DTC	61/55%	60.74%	60.25%	57.60%
KNN	28.49%	20.50%	26.25%	20.08%
ADB	52.68%	54.94%	52.43%	53.23%
LDA	54.30%	53.11%	53.92%	53.16%
CNN	97.04%	96.96%	97.02%	96.83%

Table 1 presents a detailed performance comparison of the traditional machine learning classifiers such as SVM, DTC, KNN, ADB, and LDA against the proposed CNN model for baby cry classification. Among the existing methods, KNN exhibits the weakest performance with an accuracy of 28.49%, indicating poor discrimination of overlapping MFCC features, while SVM also shows limited effectiveness with an accuracy of 46.50% and relatively low recall. DTC, ADB, and LDA demonstrate moderate performance, achieving accuracies in the range of 52–61%, suggesting that tree-based, boosting, and linear methods can partially capture cry-specific patterns but struggle with complex

acoustic variations. In contrast, the proposed CNN model significantly outperforms all baseline algorithms, achieving an accuracy of 97.04%, precision of 96.96%, recall of 97.02%, and F-score of 96.83%. This substantial improvement highlights the CNN's superior ability to learn discriminative temporal and spectral representations from MFCC features, thereby providing highly reliable and robust baby cry classification suitable for early health monitoring applications.

5. CONCLUSION

The study presents a CNN-based system for baby cry classification aimed at early identification of infant conditions, highlighting the capability of DL techniques in processing complex audio signals. The framework incorporates a complete pipeline including dataset handling, MFCC-based feature extraction, preprocessing, evaluation of ML baseline models, and a CNN model integrated within a user-friendly GUI. Results show that ML models such as SVM, KNN, DT, ADB, and LDA exhibit lower performance due to the nonlinear and overlapping nature of cry patterns. In contrast, the CNN model achieves superior accuracy, precision, recall, and F-score by effectively capturing deep temporal and spectral features. Evaluation metrics such as CM and ROC curves validate the consistency and reliability of the predictions. The system successfully classifies unseen audio samples into categories like hunger, discomfort, burping, tiredness, and belly pain, demonstrating practical applicability.

REFERENCES

[1] Lahti, K.; Vänskä, M.; Qouta, S.R.; Diab, M.Y.; Perko, K.; Punamäki, R.L. Maternal experience of their infants' crying in the context of war trauma: Determinants and consequences. Infant Ment. Health J. 2019, 40, 717–734.

[2] Halpern, R.; Coelho, R. Excessive crying in infants. J. Pediatr. 2016, 92, S40–S45.

[3] Sharma, K.; Gupta, C.; Gupta, S. Infant weeping calls decoder using statistical feature extraction and Gaussian mixture models. In Proceedings of the 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kanpur, India, 6–8 July 2019; pp. 1–6.

[4] Maghfira, T.N.; Basaruddin, T.; Krisnadhi, A. Infant cry classification using CNN-RNN. J. Phys. Conf. Ser. 2020, 1528, 012019.

[5] Franti, E.; Ispas, I.; Dascalu, M. Testing the Universal Baby Language hypothesis - automatic infant speech recognition with CNNs. In Proceedings of the 2018 41st International Conference on Telecommunications and Signal Processing (TSP), Athens, Greece, 4–6 July 2018; pp. 1–4.

[6] Li F, Cui C, Hu Y. Classification of Infant Crying Sounds Using SE-ResNet-Transformer. *Sensors*. 2024; 24(20):6575. https://doi.org/10.3390/s24206575

[7] Mekhfioui M, Fadel W, Hammouch FE, Laayati O, Bouchouirbat M, El Bazi N, Satif A, Boujiha T, Chebak A. Development of a Baby Cry Identification System Using a Raspberry Pi-Based Embedded System and Machine Learning. *Technologies*. 2025; 13(4):130. https://doi.org/10.3390/technologies13040130

[8] Zayed Y, Hasasneh A, Tadj C. Infant Cry Signal Diagnostic System Using Deep Learning and Fused Features. *Diagnostics*. 2023; 13(12):2107. https://doi.org/10.3390/diagnostics13122107

[9] Liang Y-C, Wijaya I, Yang M-T, Cuevas Juarez JR, Chang H-T. Deep Learning for Infant Cry Recognition. *International Journal of Environmental Research and Public Health*. 2022; 19(10):6311. https://doi.org/10.3390/ijerph19106311

[10] Herlea DM, Iancu B, Ardelean E-R. A Study of Deep Learning Models for Audio Classification of Infant Crying in a Baby Monitoring System. *Informatics*. 2025; 12(2):50. https://doi.org/10.3390/informatics12020050

[11] Khalilzad Z, Hasasneh A, Tadj C. Newborn Cry-Based Diagnostic System to Distinguish between Sepsis and Respiratory Distress Syndrome Using Combined Acoustic Features. *Diagnostics*. 2022; 12(11):2802. https://doi.org/10.3390/diagnostics12112802

[12] Shayegh SV, Tadj C. Deep Audio Features and Self-Supervised Learning for Early Diagnosis of Neonatal Diseases: Sepsis and Respiratory Distress Syndrome Classification from Infant Cry Signals. *Electronics*. 2025; 14(2):248. https://doi.org/10.3390/electronics14020248

[13] Tsalera E, Papadakis A, Samarakou M. Comparison of Pre-Trained CNNs for Audio Classification Using Transfer Learning. *Journal of Sensor and Actuator Networks*. 2021; 10(4):72. https://doi.org/10.3390/jsan10040072

[14] Ozcan T, Gungor H. Baby Cry Classification Using Structure-Tuned Artificial Neural Networks with Data Augmentation and MFCC Features. *Applied Sciences*. 2025; 15(5):2648. https://doi.org/10.3390/app15052648

[15] Khalilzad Z, Kheddache Y, Tadj C. An Entropy-Based Architecture for Detection of Sepsis in Newborn Cry Diagnostic Systems. *Entropy*. 2022; 24(9):1194. https://doi.org/10.3390/e24091194